DOCUMENT RESUME

ED 300 386 TM 012 201

AUTHOR Perkins, Kyle
TITLE The Respective Advantages and Disadvantages of Different Ways of Measuring the Instructional Sensitivity of Reading Comprehension Test Items.
PUB DATE Mar 87
NOTE 20p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Bayesian Statistics; Correlation; Elementary Secondary Education; Evaluation Methods; Experiments; *Factor Analysis; *Pretests Posttests; Reading Comprehension; *Reading Tests; *Research Methodology; *Test Items; Test Validity
IDENTIFIERS *Factorial Designs; *Instructional Sensitivity; Instructional Variables

ABSTRACT

In this paper four classes of procedures for measuring the instructional sensitivity of reading comprehension test items are reviewed. True experimental designs are not recommended because some of the most important reading comprehension variables do not lend themselves to experimental manipulation. "Ex post facto" factorial designs are recommended because they allow the investigator to select and to study two or more variables and their interactions simultaneously. Correlational studies are recommended for the establishment of relationships between variables in the planning stages of an experimental study. Of the large family of instructional sensitivity indices reviewed, only one, the pre-to-post difference index, is recommended because of its demonstrated robustness and simplicity. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED300386

The Respective Advantages and Disadvantages of Different Ways of Measuring the Instructional Sensitivity of Reading Comprehension Test Items

Kyle Perkins
Department of Linguistics
Southern Illinois University
Carbondale, Illinois 62901-4517
U.S.A.

(Measuring Instructional Sensitivity)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Kyle Perkins

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 012 201

ABSTRACT

In this paper four classes of procedures for measuring the instructional sensitivity of reading comprehension test items are reviewed. True experimental designs are not recommended because some of the most important reading comprehension variables do not lend themselves to experimental manipulation. Ex post facto factorial designs are recommended because they allow the investigator to select and to study two or more variables and their interactions simultaneously. Correlational studies are recommended for the establishment of relationships between variables in the planning stages of an experimental study. Of the large family of instructional sensitivity indices reviewed only one, the pre-to-post difference index, is recommended because of its demonstrated robustness and simplicity.

INTRODUCTION

The purpose of this paper is to introduce four different classes of procedures for measuring the instructional sensitivity of reading comprehension test items and to highlight the advantages and disadvantages of each class of procedures. The four classes of procedures to be discussed are (1) true experimental designs; (2) ex post facto factorial designs; (3) correlational studies; and (4) instructional sensitivity indices. A principal thesis of this paper is that true experimental designs are not particularly suitable for measuring the instructional sensitivity of reading comprehension test items, but the remaining three classes of procedures are suitable.

Instructional sensitivity

The rationale for estimating and measuring instructional sensitivity is based on the proposition that test results, especially criterion-referenced test results, are obliged to reflect the effects of instruction on examinee learning and achievement. The conventional wisdom is that uninstructed examinees will exhibit a lower level of performance on a test than instructed examinees, and the difference, or gain, from preinstruction to postinstruction, in test scores should provide an estimate of instructional effectiveness. Consequently, there is a perceived need for procedures which determine whether test items are sensitive to instruction, that is, instructional sensitivity.

Haladyna and Roid (1981) provided the following definition of instructional sensitivity:

> Instructional sensitivity is defined as the tendency for an item to vary in difficulty as a function of instruction. When instruction is

> reasonably effective, items given to unistructed students should appear difficult and the very same items when administered to instructed students should appear easy. Items that do not detect differences from pretest to posttest occasions should be reviewed to ascertain if (a) instruction has been faulty or (b) the item is flawed in some way or (c) the item is inappropriate and does not belong in the domain. Instructional sensitivity is therefore a test item concept which reflects the role played by instruction in affecting examinee performances on individual test items (p. 40).

True experimental designs

In general, true experimental designs can be characterized as follows. The researcher manipulates the independent variable and studies the effects of the independent variable on the dependent variable. The independent variable is so named because the researcher directly changes it, and, therefore, its value is independently manipulated. The dependent variable is so named because its value is dependent on the value of the independent variable. The dependent variable is also the criterion which determines whether anything of importance happens when the independent variable is manipulated.

A true experimental design always has three essential attributes: (1) there are one or more control groups; (2) there are pretest data available to indicate whether the groups were similar or different in the trait to be studied; and (3) there is randomized selection and assignment of examinees to the groups.

The example of a true experimental design to be offered here is the pretest-posttest control group, a commonly utilized design. At first blush, this design, which takes the following form, seems to be ideally suited for measuring the effectiveness of reading comprehension instruction:

R	O1	X	O2
R	O3		O4

R refers to the randomization of the control and experimental groups of examinees; X refers to the experimental treatment (in this context, the instruction); O1 is the pretest for the experimental group; O2 is the posttest for the experimental group; O3 is the pretest for the control group; O4 is the posttest for the control group.

A true experimental design like the pretest-posttest control group permits the researcher to control for a number of factors that may influence or cloud the experimental findings (provide rival hypotheses) and seriously reduce the usefulness and generalizability of the data. Several such factors are discussed below.

History refers to events that occur between the first observation and the second in addition to the experimental treatment that may influence examinee behavior. A true experimental design controls for history because any events that may have produced an O1-O2 change will also produce an O3-O4 change as well. Unique intrasession history must be controlled for by a randomization of experimental occasions with restrictions to achieve balanced representation of likely sources of bias. If all the control examinees are treated in a single session and the experimental examinees are treated in another session, then irrelevant unique events in either session become rival hypotheses explaining the O1-O2 versus the O3-O4 difference.

Statistical regression refers to regression toward the mean when examinees who score at the extremes of the normal distribution are included in the experimental treatment. If, for example, only the poorest readers were chosen to receive instruction, then they would have low scores on the pretest and would likely have higher scores on the posttest. The important point here is that this dependable result will not necessarily be due to the effects of the instruction, but will be a tautological consequence of the imperfect correlation between the pre and posttest scores. Regression is controlled for by random

assignment to the experimental and control groups; therefore, each group should be expected to regress to the same extent.

Selection bias results from differential selection of subjects for the experimental and control groups. This can be controlled for by appropriate randomization procedures and careful pretesting of the examinees.

Interaction of testing and the experimental treatment is controlled for by using regular classroom examinations as the experimental observations.

Maturation, instrumentation, and experimental mortality are additional competing hypotheses which are controlled for with this design. The use of a true experimental design is highly appropriate for this research because it permits the researcher to control for internal and external validity and to generalize the findings to other situations (Campbell and Stanley, 1963).

In sum, a true experimental design controls for the following sources of error: (1) experimental errors due to change in the dependent variable being produced by any factor other than the independent variable; (2) random errors which sometimes favor one experimental condition and sometimes the other on a chance basis, given repetition of the basic experiment; (3) constant errors which favor the same experimental condition given repetitions of the basic experiment.

In many published studies which have employed true experimental designs of this sort, researchers simply compute gain scores for the experimental and control groups and then calculate t-statistics between the scores. A more appropriate analysis should be employed: randomized blocking on pretest scores with analysis of covariance. Analysis of covariance permits the researcher to increase the precision of randomized experiments by partitioning out variance in the dependent variable accounted for by the concomitant variables. In this way the influence of the independent variable (instruction) can be more accurately assessed (Wildt and Ahtola, 1978).

The preceding section has outlined the merits of a true experimental design for estimating the instructional sensitivity of reading comprehension test items. In addition, the researcher would also insure that (1) the objectives and skills of the instructional program would be completely stated in advance of the study; (2) all the objectives and skills would be taught in the instructional program; and (3) all the skills and objectives would be assessed during the pre and posttesting. In other words, the true experimental design researcher would make certain in advance that there was goodness of fit among the goals and objectives of the course, the instructional program, and the assessment program.

Thus far, a true experimental design appears to be ideally suited for testing the effectiveness of instruction in reading comprehension; however, there are two major flaws which should preclude its use: (1) an ethical flaw because the control group does not receive instruction and (2) such a design does not allow the researcher to study clusters of variables simultaneously to develop a more meaningful description of the reading process (Kamil, 1984).

The reasons for rejecting a true experimental design are indirectly reflected in Kamil's (1984) discussion of the inherent limitations on reading research per se:

> Reading research is also shaped by the environment in which it is conducted. The choices are often polarized as being in the field or in the laboratory. It is more difficult, but not impossible, to conduct experimental research in the field. In fact, descriptive data are difficult, but not impossible, to collect in a laboratory.
>
> Similarly, experimental research is easiest to conduct in a laboratory. The control of extraneous variables may produce ideal experimental conditions at the expense of ecological validity. The use of randomized procedures, replications, and other techniques, however, can compensate for the lack of control in the field.
>
> Prototypically, an experimenter manipulates variables and observes changes in performance. Many problems in reading research cannot (for many reasons, including ethical considerations) be studied this way. For example, one cannot induce reading disability to study the effectiveness of methods of remediation (p. 47).

Thus, it would appear that the use of a true experimental design to assess instructional sensitivity must be ruled out for ethical and ecological validity reasons and also for the reason that such a design cannot accommodate Kamil's proposition that clusters of variables must be researched to develop a thorough description of the reading comprehension process.

Ex post facto factorial design research

It is possible to conduct rigorous, systematic research which is not susceptible to the flaws entailed by true experimental designs by using ex post facto behavioral research. In conducting ex post facto studies, the researcher selects rather than manipulates the independent variables. With an ex post facto study the researcher uses data which are natural, unobtrusively obtained by-products of the research setting, thereby preserving the ecological validity of the readig classroom.

A factorial design allows the researcher to study the effects of two or more independent variables simultaneously. Thus, this design permits the reading researcher to incorporate different factors (or clusters in Kamil's framework) which should provide more meaningful insight into the reading comprehension process. In addition to allowing the researcher to conduct two experiments simultaneously with the same subjects thereby saving time, a factorial design permits the study of possible interactions between the effects of the independent variables. Factorial designs can be used with both experimental and ex post facto classes of behavioral research, and factorial designs have traditionally been analyzed using the analysis of variance.

An example of how one might use an ex post facto factorial design in reading research is illustrated below. Suppose that a researcher decided to investigate the effects of two independent variables: an instructional factor

such as time on task and an environmental factor such as teacher-student interaction patterns. The dependent variable might be gain scores on the reading skills and objectives that were taught in the instructional period and assessed by the pre and posttests.

The following diagram illustrates such a factorial design:

		Teacher-student interaction patterns (Factor B)		
		B1	B2	Effect of time on task
Time on task (Factor A)	A1	$\bar{X}$ A1B1	$\bar{X}$ A1B2	$\bar{X}$ A1
	A2	$\bar{X}$ A2B1	$\bar{X}$ A2B2	$\bar{X}$ A2
Effect of teacher-student interaction patterns		$\bar{X}$ B1	$\bar{X}$ B2	

With the mean gain scores as the dependent variable and with a two-way analysis of variance it is possible to determine whether there was a significant difference in the gain scores according to (a) time on task (Factor A); (b) teacher-student interaction patterns (Factor B); and (c) whether time on task produced a larger gain when there was one teacher-student interaction pattern versus the other (A by B). In sum, ex post facto factorial designs offer the reading researcher more advantages than true experimental designs.

Correlational research

Because correlational studies do not entail direct manipulation of the variables by the researcher, they, at first blush, appear to be desirable for reading research in that the researcher does not disturb the ecological validity of the classroom by the obtrusive manipulation of the independent variables.

In conducting a correlational study the researcher seeks to determine if there is a relationship between pairs of variables under investigation; however, the researcher cannot determine if there is a causal relationship between two variables because significant correlation does not imply causation.

Published correlational research has produced rather mixed results. Rosenshine and Stevens (1984) reviewed a number of studies whose focus was the relationship between content covered in reading classes and student achievement as measured by different reading comprehension instruments. They found that the reported correlations ranged from .2 to .7. They also commented on a common flaw in instructional sensitivity studies:

> It may be difficult to obtain higher correlations because there is not a perfect correspondence between the content given in a class and the content measured on the test and also because content covered is not identical with content mastered (p. 774).

As we have previously mentioned, it is absolutely vital that there be goodness of fit between content taught and content tested for all instructional sensitivity studies, including correlational studies. The positive aspect of correlational studies is that they are not ecologically obtrusive because the variables are not directly manipulated; on the other hand, the weakness of correlational studies is that they do not allow the researcher to infer causal relationships. As a result, correlational studies may have limited usefulness in the pre-experimental stages of research when the investigator is forming plausible hypotheses to be experimentally confirmed or disconfirmed.

Potential independent variables

According to Kamil (1984) there are three major classes of variables to be investigated in reading research: (1) instructional variables; (2) reader/

learner factors; and (3) environmental factors. Other rapidly developing themes include text organization, schema theory, story grammars, metacognition, a revitalization of eye movement studies, and the effect of teaching writing/reading on reading/writing.

Among the most interesting instructional variables is time on task. Rosenshine and Stevens (1984) listed three important time on task variables:

> (a) the average engagement rate of the students (i.e., the percentage of allocated time that the students were engaged in academic activities), (b) the minutes per day that the students were academically engaged, and (c) student engagement rates in different settings (p. 782).

Variables which come under the generic label "content covered" include the number of basal texts completed (Anderson, Evertson, and Brophy, 1979); grade level achievement in a computer-assisted mastery-learning program (Paulsen and Macken, 1978); the number of books that students have read (Harris, Morrison, Serwer, and Gold, 1968); the percentage of students who were offered the opportunity to learn each test item (Chang and Raths, 1971); and textbook content (Chall, 1977). The purpose in listing these variables is to suggest different factors which may usefully be incorporated into ex post facto factorial design reading research on the effectiveness of instruction in reading comprehension classes.

Instructional sensitivity indices

In the event that a reading teacher has interest in estimating the instructional sensitivity of reading comprehension instruction, but for a variety of reasons including administrative procedures cannot conduct factorial studies, it is possible for reading teachers to utilize various instructional sensitivity

indices using data from their own classes. The only requirements for these indices are that the test be administered twice as a pre and postinstruction measure and that the test assess the skills objectives which were taught in class.

There has been considerable use made of the following instructional sensitivity indices: (1) the pre-to-post difference index (PPDI) introduced by Cox and Vargas (1966); (2) the percent of possible gain (PPG) introduced by Brennan and Stolurow (1971); (3) the combined-samples (instructed and uninstructed students) point-biserial correlation (COMPBI) introduced by Haladyna (1974); (4) four indices, O1, 11, 00, and 10, introduced by Popham (1971); (5) three item discrimination indices, D1, D2, and D3, suggested by Helmstadter (1974); (6) the normalized difference in Rasch one-parameter item response theory item difficulty estimates for the uninstructed and instructed samples introduced by Haladyna and Roid (1981); and (7) three indices based on the Bayes' theorem proposed by Helmstadter (1974) and Haladyna and Roid.

PPDI is the difference in item difficulty (determined as the proportion of correct responses) observed when an item is presented to uninstructed examinees, and then after they have been instructed. According to Roid and Haladyna (1982), "this index is the simple difference between the difficulty observed in the post-instruction group and the difficulty observed in the pre-instruction group. It therefore ranges from -1.00 to +1.00, with indexes of zero or lower quite rare. In most instructional settings, this index typically ranges from .10 to .60" (pp. 218-219).

PPG is a variation of PPDI, and it is computed by the following formula:

$$PPG = \frac{\text{difficulty (posttest)} - \text{difficulty (pretest)}}{1 - \text{difficulty (pretest)}}$$

PPG "is based on the notion that with any pretest item difficulty, there can only be a maximum gain, which is 1.00 minus the pretest difficulty" (Roid and Haladyna, 1982, p. 219). This quote does not address the possibility that the posttest difficulty might be higher than the pretest difficulty, a phenomenon that can occur in English as a Second Language measurement.

The point-biserial correlation is a common procedure for computing item discriminability with point-biserial correlations of .40 and higher being acceptable discrimination indices. Haladyna (1974) recommended combining uninstructed and instructed samples for the point-biserial correlation because more test-score variance is available with the combined samples, and COMPBI is less affected by range restriction. Haladyna noted that the COMPBI seems to function as a type of instructional sensitivity index.

The 01, 11, 00, and 10 indices were introduced by Popham (1971). When examinees take a pretest and a posttest, there are four logical states of affairs: (1) 01, an incorrect response on the pretest and a correct response on the posttest; (2) 11, correct responses on both the pre and posttest; (3) 00, incorrect responses on both tests; and (4) 10, a correct response on the pretest and an incorrect response on the posttest. The percentage in the 01 cell for examinees who took both tests would provide an index of instructional sensitivity.

D1, D2, and D3 are three separate variants of a classical item discrimination index which is computed using the sample separation technique. These variants involve defining "high knowledge" and "low knowledge" groups in different ways. The following descriptions came from Helmstadter (1974).

D1 assumed that persons in the top one-third of the class on the posttest were "high knowledge" and persons in the bottom one-third of the class on the posttest were "low knowledge".

D2 involved combining scores from the pre and posttest as if they constituted one large class and then assuming that persons in the top one-third of

this doubled class were "high knowledge" and that persons in the botton one-third of this class were "low knowledge".

D3 assumed that the pretest scores represented a "low knowledge" group and that posttest scores represented a "high knowledge" group.

The application of the Rasch one-parameter item response model can produce item characteristics which can be used for instructional sensitivity research. Haladyna and Roid (1981) used the normalized equivalent of the pre-to-posttest difference in logits of difficulty (ZDIFF) and found a high degree of correspondence between ZDIFF and other instructional sensitivity indices.

Helmstadter (1974) and Haladyna and Roid (1981) proposed three instructional sensitivity indices which were based on the Bayes' theorem (Bayes, 1763). Bayes, an English clergyman who died in 1760, produced a formal derivation of the equality among certain probabilities, and the theorem "is based on the fact that the joint probability of two events A and B can be written as the probabililty of one of the events and the conditional probability of the second events, given the first event" (Iversen, 1984, p. 12). The Bayes' theorem relates conditional and marginal probabilities as follows:

$$P(B|A) = P(A|B)\ P(B)/P(A).$$

The three indices, based on the Bayes' theorem and proposed by Haladyna and Roid, require collateral information; the most appropriate type of information is preinstruction test results. The three Bayesian indices are (a) B1, the probability that an examinee knows the content material given that the correct response was selected; (b) B2, the probability that an examinee does not know the content material given that the incorrect response was selected; and (c) B3, the probability that a correct decision will be made about the examinee's knowledge of the content given the results of performance on that item, that

is mastery or nonmastery. An implication is that items with high B3 indices augment the test administrator's probability of correctly identifying mastery and nonmastery students. The computing formulae for the Bayesian indices are quite complex and will not be reproduced here.

Thomas Haladyna (personal communication) has noted that values of .70 and higher for the Bayesian indices reflect reasonable instructional sensitivity. This recommendation comes from a consideration of reasonable pre-to-post difference indices and converting them to the equivalent Bayesian indices.

Perkins (1987) reported a comparative analysis of all these indices using pre and postinstruction data from a standardized grammar examination; the subjects were English as a Second Language students enrolled in an intensive English grammar class. After the indices were calculated for the forty items in the test, they were submitted to Pearson product-moment correlation. The most robust indices were ZDIFF and PPDI; robust in this context refers to the magnitude of the correlations and the number of significant correlations entered.

There were many vexing problems with the Bayesian indices: (1) low pretest item difficulties generated spuriously high B1 indices; (2) low pretest and posttest item difficulties generated spurious B2 indices; and (3) the B3 index was influenced by ceiling and floor effects.

The practical implication for classroom reading teachers from Perkins' (1987) research is that of all the indices studied, the PPDI is the most highly recommended to classroom teachers for the following reasons: (1) it is the easiest index to calculate; (2) it was consistently related to many of the other indices; and (3) its robustness has been demonstrated in first language, non-language content settings (Herbig, 1976).

SUMMARY

This paper offers the follow principal recommendations to reading researchers and instructors: (1) instructional sensitivity is an important test item concept in reading instruction and assessment because it can serve as a signal to ascertain if an item is faulty or belongs to another domain or if instruction has been faulty; (2) true experimental designs are not always practical for reading instructional sensitivity studies because they may entail a disturbance and an intrusion into the ecological validity of the reading classroom; (3) if the researcher has access to a setting and samples of subjects, ex post facto factorial designs are recommended because they permit the researcher to select and to study the effects of two or more independent variables simultaneously and to study the possible interactions between the effects of the independent variables; (4) correlational studies are recommended for the pre-experimental stages of research when the investigator is forming plausible hypotheses to be tested; correlational studies establish a relationship between variables, but they do not allow one to infer causal relationships; and finally, (5) if teachers cannont conduct the previous types of studies, they may elect to utilize instructional sensitivity indices; of the lot reviewed in this report, PPDI is the only index recommended for use because of its robustness and simplicity.

REFERENCES

Anderson, R., Evertson, C., and Brophy, J. (1979) An experimental study of effective teaching in first grade groups. Elementary School Journal, 79, 132-222.

Bayes, T. (1763) Essay towards solving a problem in the doctrine of chances. Philosophical Transactions, 53, 370-418.

Brennan, R. L., and Stolurow, L. M. (1971) An empirical decision process for formative evaluation. Research Memorandum No. 4. Cambridge, MA: Harvard University CAI Laboratory.

Campbell, D. T., and Stanley, J. C. (1963) Experimental and Quasi-experimental Designs for Research. Chicago: Rand McNally.

Chall, J. S. (1977) An Analysis of Textbooks in Relation to Declining SAT Scores (A Report to the Advisory Panel on Score Decline sponsored by the College Entrance Examination Board) March.

Chang, S. S., and Raths, J. (1971) The school's contribution to the cumulating deficit. Journal of Educational Research, 64, 272-276.

Cox, R. C., and Vargas, J. (1966) A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, April.

Haladyna, T. (1974) Effects of different samples on item and test characteristics

of criterion-referenced tests. Journal of Educational Measurement, 11, 93-100.

Haladyna, T., and Roid, G. (1981) The role of instructional sensitivity in the empirical review of criterion-referenced test items. Journal of Educational Measurement, 18, 39-52.

Harris, A. J., Morrison, C., Serwer, B. L., and Gold, L. (1968) A continuation of the CRAFT project. New York: City University of New York.

Helmstadter, G. C. (1974) A comparison of Bayesian and traditional indexes of test item effectiveness. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, April.

Herbig, M. (1976) Item analysis by use in the pretests and posttests: A comparison of different coefficients. Programmed Learning and Educational Technology, 13, 49-54.

Iversen, G. R. (1984) Bayesian Statistical Inference. (Sage University paper No. 43. Series: Quantitative Applications in the Social Sciences). Beverly Hills, CA and London: Sage Publications, Inc.

Kamil, M. L. (1984) Current traditions of reading research. In P. D. Pearson (ed) Handbook of Reading Research. New York and London: Longman, 39-62.

Paulsen, J., and Macken, M. (1978) Report on the level of achievement in CAI reading programs. Palo Alto, CA: Computer Curriculum Corporation.

Perkins, K. (1987) A comparison of instructional sensitivity indices. Paper presented at the TESL Canada Summer Forum, Concordia University, Montreal, July 25.

Popham, W. J. (1971) Indices of adequacy for criterion-referenced tests. In W. J. Popham (ed) Criterion-referenced Measurement. Englewood Cliffs, NJ: Educational Technology Publications.

Roid, G., and Haladyna, T. (1982) A Technology for Test-item Writing. New York; Academic Press.

Rosenshine, B., and Stevens, R. (1984) Classroom instruction in reading. In P. D. Pearson (ed) Handbook of Reading Research. New York and London: Longman, 745-798.

Wildt, A. R., and Ahtola, O. T. (1978) Analysis of Covariance. (Sage University Paper series on Quantitative Applications in the Social Sciences, 07-012. Beverly Hills, CA and London: Sage Publications, Inc.